Conference on Applied Statistics in Agriculture 2008 - 20th Annual Conference Proceedings

# DYNAMIC CLUSTERING OF CELL-CYCLE MICROARRAY DATA

Lingling An

R. W. Doerge



## Recommended Citation

An, Lingling and Doerge, R. W. (2008). "DYNAMIC CLUSTERING OF CELL-CYCLE MICROARRAY DATA," *Conference on Applied Statistics in Agriculture*. https://doi.org/10.4148/2475-7772.1094

# DYNAMIC CLUSTERING OF CELL-CYCLE MICROARRAY DATA

Lingling An and R.W. Doerge
Department of Statistics
Purdue University
150 North University Street
West Lafayette, IN 47907 USA

## Abstract

The cell cycle is a crucial series of events that are repeated over time, allowing the cell to grow, duplicate, and split. Cell-cycle systems play an important role in cancer and other biological processes. Using gene expression data gained from microarray technology it is possible to group or cluster genes that are involved in the cell-cycle for the purpose of exploring their functional co-regulation. Typically, the goal of clustering methods as applied to gene expression data is to place genes with similar expression patterns or profiles into the same group or cluster for the purpose of inferring the function of unknown genes that cluster with genes of known function. Since a gene may be involved in more than one biological process at any one time, co-regulated genes may not have visually similar expression patterns. Furthermore, the time duration for genes in a biological process may differ, and the number of co-regulated patterns or biological processes shared by two genes may be unknown. Based on this reasoning, biologically realistic gene clusters gained from gene co-regulation may not be accurately identified using traditional clustering methods. By taking advantage of techniques and theories from signal processing, it possible to cluster cell-cycle gene expression profiles using a dynamic perspective under the assumption that different spectral frequencies characterize different biological processes.

***Keywords:*** clustering, dynamic, cell cycle, gene expression

## 1. Introduction

In biology, a cell cycle is a process where the genetic contents in cells are replicated so that the cells propagate. Cell cycles contain four main phases (Wu and Bonner 1981): S (synthesis), M-phase (mitosis), G1 (the first gap) and G2 (the second gap). In S phase, the genome is duplicated and in M phase, the nucleus is divided. G1 refers to the first gap phase from the end of the previous M phase to the beginning of the next S phase, and G2 refers to the second gap phase from the end of the previous S phase to the beginning of the next M phase.

The levels of mRNA for some genes oscillate as a function of the cell cycle (Hereford et al. 1981; Matsumoto et al. 1985). Recently, genome-wide studies have shown that hundreds or thousands of genes are periodically transcribed during the cell cycles of different species, including human (Whitfield et al. 1998), yeast (Cho et al. 1998; Spellman et al. 1998), and Arabidopsis thaliana (Menges et al. 2003). Investigating cell-cycle related genes can help us understand their functions, identify their roles in the cell cycle, and understand the underlying mechanisms of transcriptional (from DNA to RNA) control. Since genes with similar biological function (e.g., cellular related function) tend to share similar patterns of expression (Holter et al. 2000), one goal is to group genes according to gene expression patterns. The main purpose of

grouping or clustering gene patterns is to learn about genes of unknown function. Specifically, if genes of unknown function are clustered with genes of known function, the hope is that the function of the unknown genes can be inferred and complex biological processes better understood.

A variety of clustering methods have been proposed or employed to cluster cell cycle gene expression data. In early studies, each profile was treated as a vector of independent elements and traditional clustering approaches were employed for the analysis (e.g., Eisen et al. 1998). When the nature of time dependence is considered in the clustering procedure high quality clusters (Tai 2005) can be gained. Toward this end there are clustering methods that account for the inherent dependence of temporal data (Bar-Joseph 2004; Moller-Levet et al. 2003), such as spline-model clustering (Luan and Li 2003; Ma et al. 2006), hidden Markov models clustering (Schliep et al. 2003; Ji et al. 2003), and Bayesian method for model based clustering (Ramoni et al. 2002). However, it is important to realize that gene cluster membership obtained by these methods is static since genes are clustered based on their complete profile. Although biclustering (Madeira and Oliveira 2006; Ji and Tan 2005; Zhang et al. 2005) addresses the issue of gene co-regulation in a portion of a time interval, the time durations for genes in the same bicluster must be same.

The time (or time duration) that a gene is involved in a biological process may differ between genes since they are usually involved in many finite time (biological) processes simultaneously, and they may enter and/or exit a process at different time points. Furthermore, after exiting a process the same gene may re-enter the process at a later time. To date no clustering method sufficiently incorporates all of these issues simultaneously when clustering gene expression time series data. Here, we focus on dynamic clustering that acknowledges the non-constant nature of gene activity as their expression initiates and stops during any biologically complex process.

As microarray technology evolves (e.g., the cost drops and the quality of the data improves), researchers are designing longer time-course (i.e., 40, 100, 300 time points) gene expression experiments. Predictably, as the length of a time series increases, the information about gene expression in the cell cycle, and the relationships between and among genes, also increases. Based on this growing base of knowledge, the full expanse of gene expression information can be studied in greater detail than is currently being done by taking advantage of techniques and theories from signal decomposition (Dougherty et al. 2004). Here we apply a continuous wavelet transformation to decompose each complicated time series gene expression. A similarity measurement is then proposed to calculate the relationship between all pairwise components of gene profiles. Based on the results from hierarchical clustering of the components a two-step cluster validation procedure is proposed to both statistically determine the optimal number of clusters and evaluate the statistical significance of clusters.

## 2. Methods

When cell cycle data are considered, the frequency (or period) is a major property of the time series. Thus, the period (or frequency) is used to characterize clusters so that different characteristics represent different biological processes. Since gene expression time series are often complicated and each gene time series may contain several components that may last for a finite duration, a signal decomposition method is employed to understand the periodic patterns varying over time.

### 2.1 Signal decomposition

It is known that the Continuous Wavelet Transformation (CWT, Carmona et al. 1998) method has a number of benefits over other methods. Specifically, it provides results with a higher resolution than the Short Time Fourier Transformation (Qian 2002), and since it is a linear transformation on the signals, it is advantageous over the Wigner-Vile Distribution for multiple-component signals. We employ CWT on each time series to obtain both the time and frequency information. CWT is able to infer which frequency varies at what time interval by using a time frequency analysis. For a time series *s(t),* the CWT is:

$$W(b,a) = \frac{1}{a}\int s(t)\varphi(\frac{t-b}{a})dt \qquad (1)$$

where $\varphi(t)$ is the mother wavelet, *a* is the scale parameter, and *b* is the shift parameter. Here, the Morlet wavelet $\varphi(t) = e^{-\frac{t^2}{2\sigma^2}} e^{i2\pi f_0 t}$ is employed (Goupillaud et al. 1984). To determine the exact frequency values and their associated starting/ending time points in time-frequency plane, the Crazy Climber algorithm (Carmona et al. 1998), a ridge extraction method, is employed to capture multiple ridges on a single plane.

### 2.2 Similarity measurement

Measuring the closeness of objects by quantifying the similarity between them is important in clustering. Once each gene expression time series is decomposed into a set of component signals, the relationship between all pairwise expression signals can be assessed using the relationships among their decomposed components. A coherency function is employed to measure the similarity between components (Butte et al. 2001). For two signals *x* and *y*, their coherency is:

$$C_{xy}(f) = \frac{P_{xy}{}^2(f)}{P_{xx}(f)P_{yy}(f)} \qquad (2)$$

where $P_{xx}(f)$ and $P_{yy}(f)$ are the power spectral densities of these two signals and $P_{xy}(f)$ is their cross-spectral density. The coherency function, which is a function of frequency, ranges from 0 to the Nyquist frequency (Grenander 1959). For two signals with frequencies $f_x$ and $f_y$, the average of the coherency function in the interval $[f_x, f_y]$ is taken to represent the coherency similarity. An example of calculating the average coherency is shown in Figure 1. Throughout the remainder of this work, the average coherency is used and is referred to as "coherency".

It is anticipated that the coherency will increase as two spectral frequencies become closer, and decrease as they become further apart. However as seen in the top plot of Figure 2, there is a non-monotonic pattern in the frequency difference, suggesting that an alternative approach or modification (Figure 2) is necessary. To make the coherency monotonic in frequency difference, a modification is achieved by connecting the valleys of the coherency curves (whose values are non-increasing along the frequency differences) and constructing a curve that is monotonic in frequency difference. The modified or constrained coherency (called CoCo) can serve as a similarity measurement in clustering.

**2.3 Clustering methods**

Since biological processes can be represented by the spectral frequencies of time series data, it is anticipated that gene expression profiles can be grouped together if they share the same or similar spectral frequencies. Essentially, the relationships between genes are associated with component relationships that are obtained by measuring the similarity using the constrained coherence (CoCo). Because a single gene contains multiple components, relationships between genes based on components may be complex. Therefore, it is most beneficial to perform clustering on the components, and then the corresponding genes can be grouped according to components.

Microarray data contain both technical and biological error that together is generally referred to as noise. As a result, some components obtained from the decomposition may be noise and may result in disjoint clusters (i.e., hard or crisp clustering). As a consequence of the noise, components do not belong to any cluster. Many clustering methods have been proposed to accommodate the noise that is known to exist in microarray data. Among them, sequential clustering, dynamic agglomerative clustering, and hierarchical clustering (Jiang et al. 2003; Tseng and Wong 2005; Liang and Wang 2007) provide crisp clusters or single cluster membership. Even though in sequential clustering the clusters are formed and removed in a sequential manner, and the results are locally optimal, there is no clear criterion for selecting the clustering parameters. In dynamic agglomerative clustering (DAC) (Liang and Wang 2007), objects are clustered simultaneously and the scattered objects (i.e., noise) are placed into one cluster. This overcomes the local optimum problem, but six user-defined parameters are required for this approach so there is a substantial amount of user expertise that is required.

We rely on agglomerative hierarchical clustering with the Ward linkage (Ward 1963) for clustering one-dimensional noisy data using the proposed CoCo similarity. The Ward linkage finds compact and homogenous clusters by minimizing the variance of objects. However, there are two challenges that remain for hierarchical clustering, namely, determining the optimal number of clusters, and differentiating meaningful (i.e., significant) clusters from noisy clusters.

## 2.4 Clustering validation

### Determining number of clusters

There are a variety of approaches that have been proposed for determining the number of clusters. Here, the one-dimensional component frequencies are clustered. Among the validation approaches that are suitable for such low-dimensional data, a particularly well-known approach for hierarchical clustering involves finding the "elbow" of an error curve (Salvador and Chan 2004). Unfortunately, most approaches for finding the elbow (or the change point) lack statistical justification. Gap statistics have been proposed (Tibshirani et al. 2001) to determine the number of clusters with statistical justification, and have been successfully used to determine the optimal number of clusters in a variety of applications (Yeung et al. 2001). However, this approach fails for noisy data since it relies on between and within sum of squared distances.

A novel approach is proposed to statistically determine the number of clusters for noisy data by evaluating a two-dimensional graph where the x-axis represents the number of clusters and the y-axis denotes the merge distance which is used to split or merge clusters in hierarchical clustering. The null hypothesis under consideration is that all data points are random (i.e., no pattern) and belong to one cluster. The uniform distribution is used as a null reference distribution, where its range is the same as the original data (e.g., the component frequencies in our case). For time series gene expression data, $n$ is the total number of decomposed components and $x_i$'s are the component frequencies. The following steps are used to determine the number of clusters:

1) Perform hierarchical clustering with Ward linkage and CoCo similarity on the original data set ($x_i$) containing $n$ points, and obtain the merge distance $M_0=(d_2, \dots, d_n)$.
2) Randomly choose $n$ objects from the uniform distribution.
3) Perform hierarchical clustering on the data that are generated from step 2 and obtain the merge distance.
4) Repeat steps 2- 3 M times (usually $M \geq 1000$). For each possible number of clusters ($k$) the 95th percentile of M merge distances serves as a threshold, and the 95% threshold curve is constructed $M^*=(d_2^*, \dots, d_n^*)$.
5) Compare $M_0$ and $M^*$; the largest k which satisfies $d_k > d_k^*$ is the optimal number of clusters $k_0$.

Even though the data are one-dimensional, in the context of clustering components the proposed method on merge distance can be applied to high dimensional data as well.

### Determining significant clusters

The distinction between a noise cluster and statistically significant clusters can be assessed by statistically evaluating compactness and separation. The Silhouette metric (Rousseeuw 1987), a measure of tightness and separation of clusters, is used to assess the statistical significance of clusters. Similar to the procedure of determining the number of clusters, the procedure of determining the cluster significance by evaluating the cluster silhouette uses a uniform distribution on the component data set. The procedure is as follows:

1) For $k_0$ clusters from the original data, compute their silhouettes.

2) Randomly choose $n$ objects from the uniform distribution.
3) Perform hierarchical clustering on the generated data set from step 2 and obtain the silhouettes for $k_0$ clusters.
4) Repeat steps 2 -3 $M$ times (usually $M \geq 1000$) and obtain $M$ sets of $k_0$ silhouettes. For each silhouette of the original $k_0$ clusters, calculate its $P$-value from $M^* k_0$ values. Cluster significance is represented by its $P$-value.
5) Let the significance level be $\alpha$. A cluster is significant if its $P$-value $< \alpha$; otherwise, it is noise.

The two-step cluster validation is been proposed for the purpose of identifying a gene as belonging to a significant cluster based on a specific component in its decomposition. As described in the results of signal decomposition in Section 2.1, a gene may contain a set of components, with each component lasting a certain period of time. Although only the component frequencies are used in clustering, the time information of components can still be included in the clustering results. Thus, the concept of a dynamic cluster evolves naturally. To our knowledge dynamic clustering has not been proposed previously, therefore it is not possible to compare our proposed method with any of the current clustering and validation methods. We apply our novel method to a real microarray data set with the hope of gaining insight into the dynamic association among genes.

## 3. Clustering Human Cancer Cell Cycle Data

Genome-wide microarray experiments have been conducted for the purpose of understanding the complex dynamic biological processes and gene functions in cell cycles for human cancer cell lines (HeLa) (Whitfield et al. 2002). The cells were synchronized by three different methods: a double thymidine block method, a thymidine-nocodazole block method, and a mitotic shake-off. Data were collected in five independent experiments. Using the Fourier transformation, 1,134 clones (representing 874 genes) were detected to have periodic patterns in all experiments. For the remainder of this example, clones will be referred to as genes. Many other studies on the HeLa data focus on detecting periodic genes (Wichert et al. 2004; Chen 2005) or static clustering (Schliep et al. 2005). However, in these studies either a list of periodic genes is given or static clusters are obtained where the gene identifications in a given cluster are fixed over time.

The proposed dynamic clustering approach, which is designated for clustering periodic profiles, is applied to the cell-cycle human HeLa data from experiment Thy-Thy3 since the data were collected at equally spaced intervals (47 time points in 46 hours). Figure 3 provides an example of a single gene's time series gene expression pattern where the cell-cycle pattern is obvious. In work by Whitfield et al. (2002) the period of the expression profiles in Thy-Thy3 experiments is found to be 15.4 hours. Here, period is the reciprocal of spectral frequency, so it also represents the character of a time series. Since the expression values for 35 of the 1,134 genes are missing at all time points in the Thy-Thy3 experiment, only the 1,099 genes with partial missing data are analyzed.

Using the Continuous Wavelet Transformation, each signal is represented by a time-frequency plane. Figure 4 provides an example of the CWT representation of the single gene that is the example in Figure 3. The pink band denotes the signal given the frequency. After performing ridge extraction on the plane, the frequency value is determined with the starting and ending time points. Using the signal decomposition method, 1,099 time series expression profiles are decomposed into a set of 1,875 components (some time series contain more than one component). The pairwise relationship for the components is calculated using the CoCo similarity. After performing hierarchical clustering and the proposed cluster validation approach two statistically significant clusters and one noise cluster are detected (Figure 5). In Figure 5 it can be seen that 992 genes are involved in the blue cluster whose characteristic (period) is 14.6 hours; and 245 genes are involved in the red cluster with period of 62.4 hours. Interestingly, 189 genes belong to these two clusters, simultaneously. Figure 6 provides the gene expression profiles in the two statistically significant clusters. The genes are ordered by phase (that is, shift or time lag). The period of 14.6 hours is close to the period of 15.4 hours that is reported by Whitfield et al. (2002) in the original work. The second period of 62.4 hours is not identified in the previous work by Whitfield et al (2002) or in other studies. The genes in this cluster require further investigation and are a worthy effort, because genes with different periods may be discovered or new expression patterns may be found.

Since time information is associated with components from the decomposition, the dynamic property of the clusters is best conveyed through panel plots (Figure 7), where the x-axis represents time. In each panel, each segment represents a time interval of a gene in the cluster so it illustrates the starting and ending points of a gene belonging to that cluster. For example, a segment of [0, 30] in the left cluster (Figure 7) with identification number of 750 represents a gene with the same identification number which contains a component signal with the period of 14.6 hours in the time interval [0 hour, 30 hours]. Furthermore, it can be seen from Figure 7 that many segments in the blue cluster (left) appear only in a partial interval while the majority of segments in the red cluster (right) remain across the whole time interval.

The time-varying number of genes in clusters is considered as the dynamic property of the clusters and can be summarized as a function of time (Figure 8). The 95% confidence interval for the number of genes is obtained by a nonparametric bootstrap method (Efron 1993). In this example the number of genes in the cluster with period 14.6 hours apparently varies along time, while the number of genes for the cluster with period 62.4 hours does not change much (Figure 8). These results can be understood by remembering that a gene may be involved in multiple processes and may change process membership from a time to time. The set of genes in a process (or cluster) at time A may not be as exactly the same set of genes at time B. Therefore, by acknowledging the non-constant nature of gene activity through the initiation and termination of expression the dynamic processes that are present in the data can be assessed via the results of dynamic clustering. It is worth noting that dynamic cluster membership is not incorporated into any other existing clustering methods, and the multiple periods in HeLa data have not yet been investigated.

## 4. Summary

For cell-cycle temporal gene expression data, spectral frequencies in time series are used to characterize biological processes. Each time series is decomposed into a set of components so that the frequencies can be studied individually. A modified coherency function is proposed to measure the similarity between pairwise components. The components (containing both meaningful and noise components) are clustered via hierarchical clustering which is coupled with the proposed similarity. A novel two-step cluster validation approach is proposed to statistically determine the number of clusters and statistically distinguish meaningful clusters from the noise cluster. The time information associated with the components is subsequently displayed as clustering results and is the most important contribution of this research, since no current statistical methods are able to address the dynamic characteristics of gene clusters.

In conclusion, the major statistical and bioinformatic contributions of this research are the concept of a time-varying cluster, and the proposed method of cluster validation for noisy data. The end result of this research provides insight into the dynamic association among time-limited co-expressed genes that have not been discovered by other clustering approaches. The proposed method is motivated by cluster analysis of microarray experiments, and it has been performed in the framework of gene expression time series. However, it is a general method that can be applied to analyze any type of periodic phenomena in other areas.

## 5. Acknowledgements

This research is supported by a grant “Functional Genomics of Plant Polyploids” from the National Science Foundation Plant Genome (DBI 0501712 and 0733857) to RWD.

## 6. References

Akansu A., Haddad P. 2001. Multiresolution signal decomposition: transforms, subbands, wavelets. *Academic Press*.

Bar-Joseph Z. 2004. Analyzing time series gene expression data. *Bioinformatics.* 20: 2493-2503

Butte A., Bao L., Reis B., Watkins T. and Kohane I. 2001. Comparing the similarity of time-series gene expression using signal processing metrics. *Journal of Biomedical Informatics*. 34: 396-405

Carmona R., Hwang W. and Torresani B. 1998. Practical time-frequency analysis: Gabor and wavelet transforms with an implementation in S. *Academic Press.*

Chen J. 2005. Identification of significant periodic genes in microarray gene expression data. *BMC Bioinformatics.* 6:286

Cho R., Campbell M., Winzeler E., Steinmetz L., Conway A.. et al. 1998. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell.* 2:65–73

Dougherty E., Shmulevich I. and Bittner M. 2004. Genomic signal processing: the salient issues. *EURASIP Journal on Applied Signal Processing.* 1:146-153

Eisen M., Spellman P., Brown P. and Botstein D. 1998. Cluster analysis and display of genome-wide expression pattern. *Proceedings of the National Academy of Sciences.* 95(25):14863-14868.

Gardner M., Hall N., Fung E., White O., Berriman M. et al. 2002. Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature.* 419:498-511

Goupillaud P., Grossman A., and Morlet J. 1984. Cycle-octave and related transforms in seismic signal analysis. *Geoexploration.* 23:85-102

Grenander U. 1959. Probability and statistics: The Harald Cramer Volume. The Nyguist frequency is that frequency whose period is two sampling intervals. Wiley.

Hereford LM, Osley MA, LudwigTR2nd, McLaughlin CS. 1981. Cell-cycle regulation of yeast histone mRNA. *Cell.* 24:367–75

Holter N., Mitra M., Maritan A., Cieplak M., Banavar J. and Fedoroff N. 2000. Fundamental patterns underlying gene expression profiles: Simplicity from complexity. *Proceedings of National Academy of Sciences*. 97: 8409-8414

Ji X., Li-Ling J., and Sun Z. 2003. Mining gene expression data using a novel approach based on hidden Markov models. *FEBS Letters.* 542: 125-131

Ji L. and Tan K. 2005. Identifying time-lagged gene clusters using gene expression data. *Bioinformatics.* 21: 509-516

Jiang D., Pei J. and Zhang A. 2003. DHC: A density-based hierarchical clustering method for time series gene expression data. *BIBE*, 393-400.

Liang F. and Wang N.. 2007. Dynamic agglomerative clustering of gene expression profiles. *Pattern Recognition Letters*. 28:1062-1076.

Luan, Y. and Li, H. 2003. Clustering of time-course gene expression data using a mixed-effects model with B-splines. *Bioinformatics* 19(4):474-482.

Ma, P., Castillo-Davis, C., Zhong, W., and Liu, J. 2006. A data-driven clustering method for time course gene expression data. *Nucleic Acids Research* 34(4):1261-1269.

Madeira S. and Oliveira A. 2006. An efficient biclustering algorithm for finding genes with similar patterns in time-series expression data. *Proceedings of 5th Asia Pacific Bioinformatics.* 67-80

Matsumoto S, Yanagida M. 1985. Histone gene organization of fission yeast: a common upstream sequence. *EMBO J.* 4:3531–38

Menges, M., Hennig, L., Gruissem, W. & Murray, J. A. H. 2003. Genome-wide gene expression in an Arabidopsis cell suspension. *Plant Mol. Biol.* 53:423–-442

Moller-Levet C., Cho K., Yin H, and Wolkenhauer O. 2003. Clustering of gene expression time-series data. Technical Report, Department of Electrical Engineering and Electronics, University of Manchester Institute of Science and Technology, UK.

Munneke B. and Schlauch K. and Simonsen K. and Beavis W. and Doerge RW. 2005. Adding confidence to gene expression clustering. *Genetics.* 170:2003-2011

Qian S. 2002. Introduction to time frequency and wavelet transforms. Prentice Hall.

Ramoni M., Sebastiani P., and Kohane I. 2002. Cluster analysis of gene expression Dynamics. *Proceedings of National Academy of Sciences*. 99: 9121-9126

Rousseeuw P. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational Applied Mathematics. 20:53-65

S. Salvador and P. Chan. 2004. Determining the number of clusters/segments in Hierarchical clustering/segmentation algorithms. 16th IEEE international conference on tools with artificial intelligence. 576-584.

Schliep A., Schonhuth A, and Steinhoff C. 2003. Using hidden Markov models to analyze gene expression time course data. Bioinformatics. 19: i255-i263

Spellman P., Sherlock G., Zhang M., Iyer V., Anders K., et al. 1998. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* 9:3273–97

Tai Y. and Speed T. 2005. Statistical analysis of microarray time course data. In Number editor, DNA Microarrays. Taylor & Francis.

Tibshirani R., Guenther W., and Trevor H. 2001. Estimating the number of clusters in a data set via the gap statistic. Journal *of the Royal Statistical Society* 63(2):411-423.

Troyanskaya O., Cantor M., Sherlock G., Brown P., Hastie T., Tibshirani R., Botstein D., and Altman R. 2001. Missing value estimation methods for DNA microarrays. *Bioinformatics.* 17:520-525

Tseng G. and Wong W. 2005. Tight clustering: a resampling-based approach for identification stable and tight patterns in data. *Biometrics.* 61:10-16.

Ward. J. Hierarchical grouping to optimize an objective function. 1963. Journal of American Statistical Association. 58:236-244.

Whitfield M., Sherlock G., Saldanha A., Murray J., Ball C., et al. 2002. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell* 13:1977–2000

Wichert S., Folianos K., and Strimmer K. 2004. Identifying periodically expressed transcripts in microarray time series data. *Bioinformatics.* 20:5-20

Wu R. and Bonner W. 1981. Separation of basal histone synthesis from S-phase histone synthesis in dividing cells. *Cell.* 27:321-330

Yeung K., Haynor D., and Ruzzo W. 2001. Validating clustering for gene expression data. *Bioinformatics*. 17:309-318.

Zhang Y., Zha H., and Chu C. 2005. A Time-Series Biclustering Algorithm for Revealing Co-Regulated Genes. *Proceedings of IEEE International Conference on Information Technology: Coding and Computing.* 32-37

Hughes T., Marton M., Jones A., Roberts C., Stoughton R., Armour C., Bennett H., Coffey E., Dai H., He Y., Kidd M., King A., Meyer M., Slade D., Lum P., Stepaniants S., Shoemaker D., Gachotte D., Chakraburtty K., Simon J., Bard M., and Friend S. 2000. Functional discovery via a compendium of expression profiles. *Cell* 102(1):109-126.

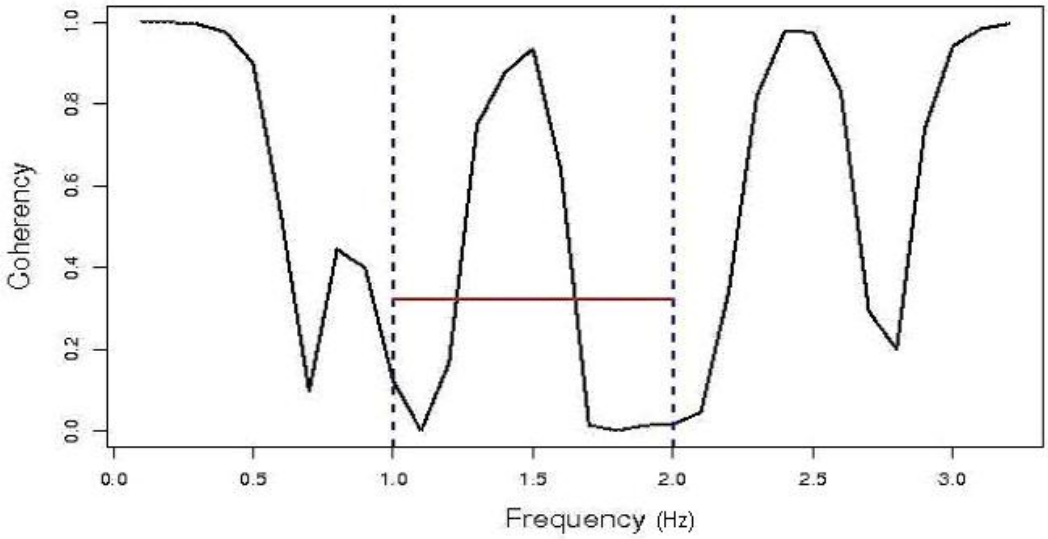
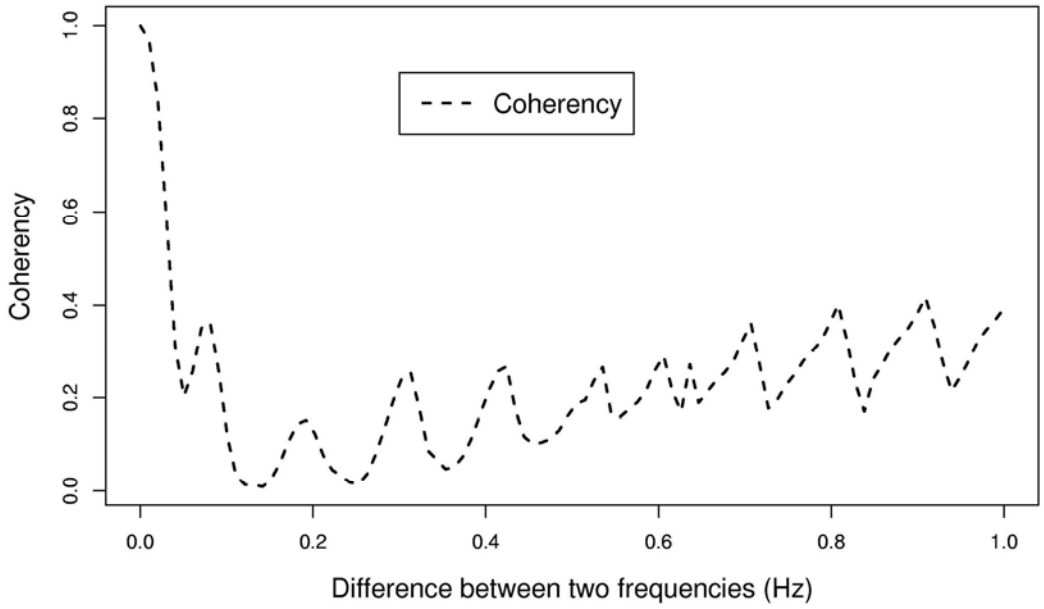



**Figure 1.** The coherency function between two signals with spectral frequencies $f_x$ =1.0 Hz and $f_y$=2.0 Hz. The average coherency between these two frequencies is 0.32.



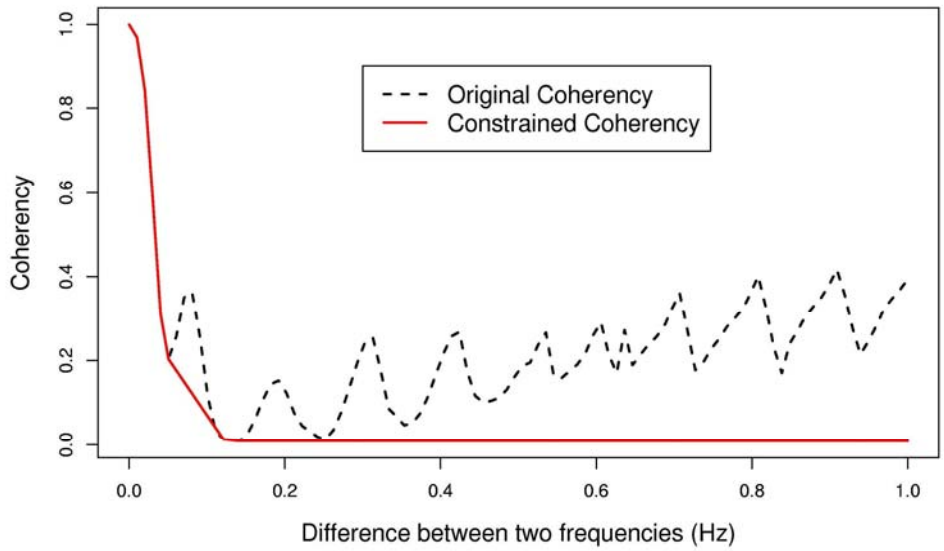




**Figure 2.** Top: A non-monotonic pattern for coherency in terms of frequency difference between two signals. Bottom: The dashed line is the coherency values with the coherency modification in red. The modified coherency is monotonic in frequency difference.

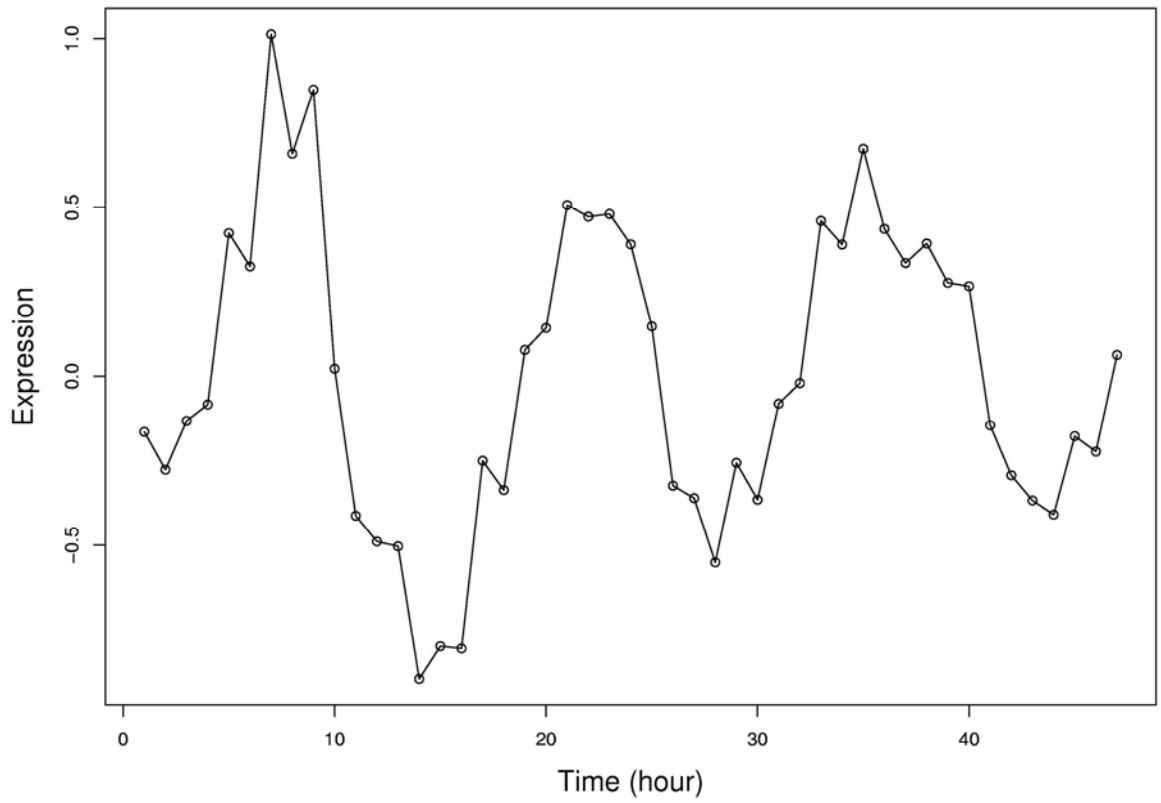



**Figure 3.** An example of time series expression for a single gene in HeLa data (Whitfield et al. 2002).

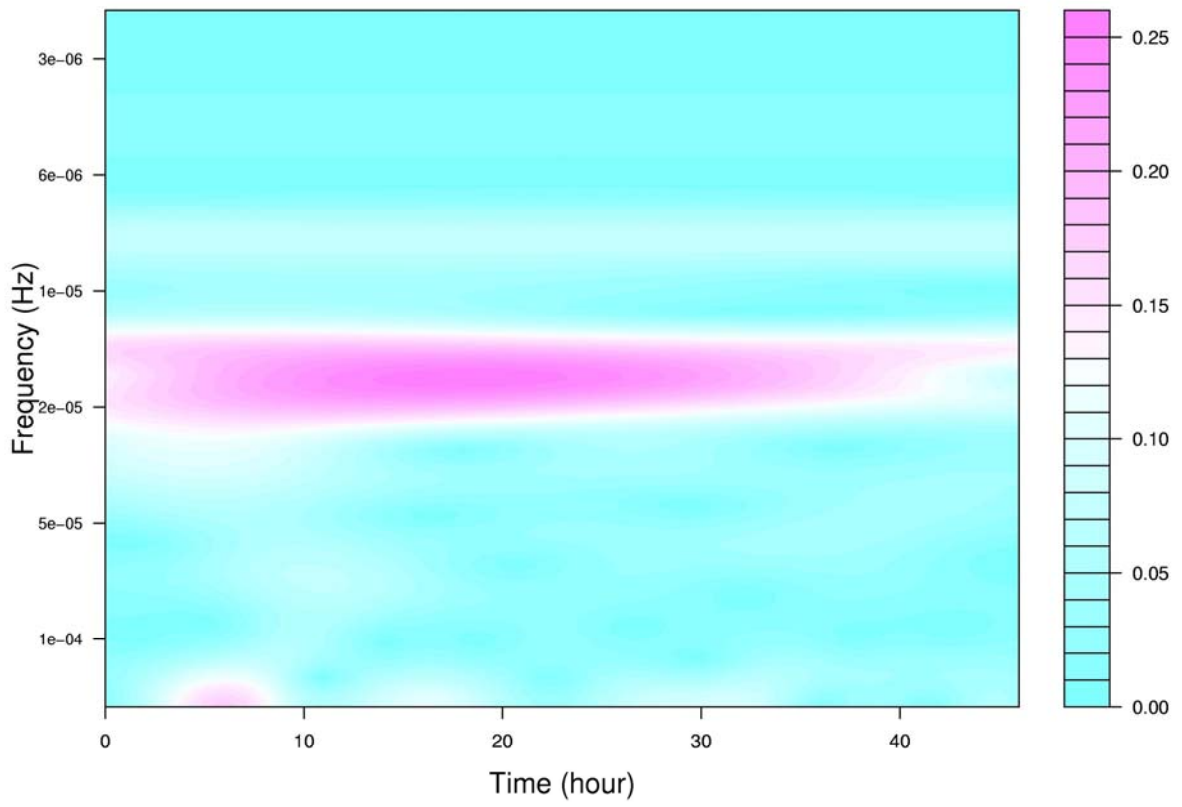



**Figure 4.** Contour plot of the modulus (i.e., |W(t,f)|) of the continuous wavelet transformation for the gene expression in Figure 3.

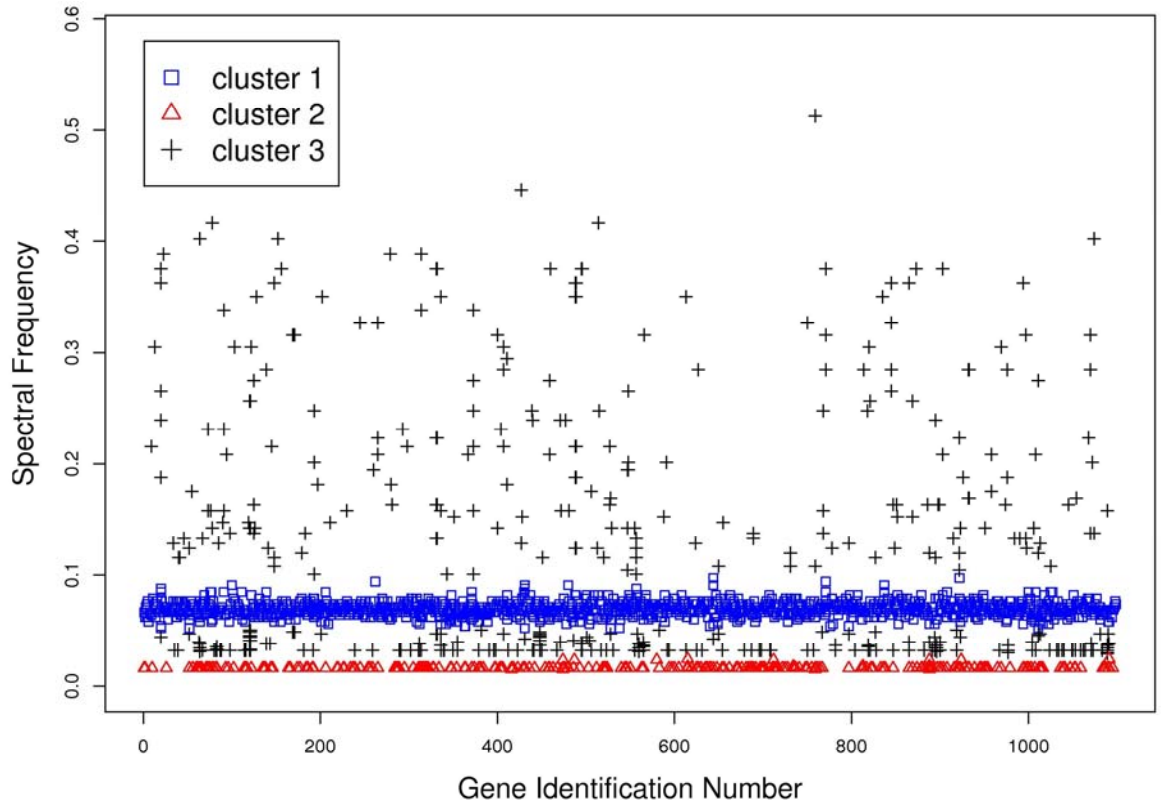



**Figure 5.** Spectral frequency (centralized) for 1,875 components in 1099 genes (Whitfield et al. 2002). Two statistically significant clusters are detected in red and blue, and the other points belong to the noise cluster, cluster 3.

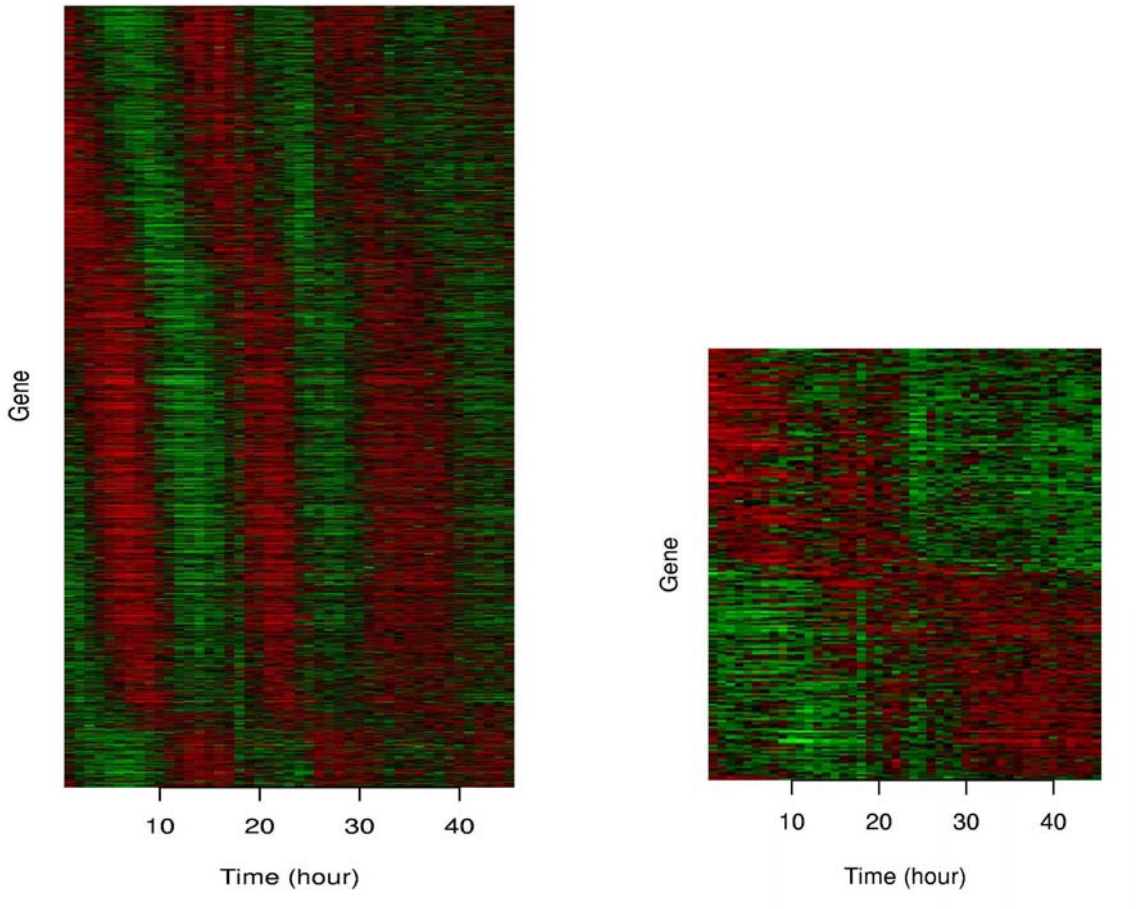



**Figure 6.** Left: expression profiles for 992 genes in the first cluster with period 14.6 hours. Right: expression profiles for 245 genes in the cluster with period 62.4 hours. The x-axis represents time (from 0 to 46 hours) and the y-axis represents genes that are ordered by their phases that are obtained in the signal decomposition. Red represents high values and green denotes low values. The expression time series data are from Whitfield et al. (2002).

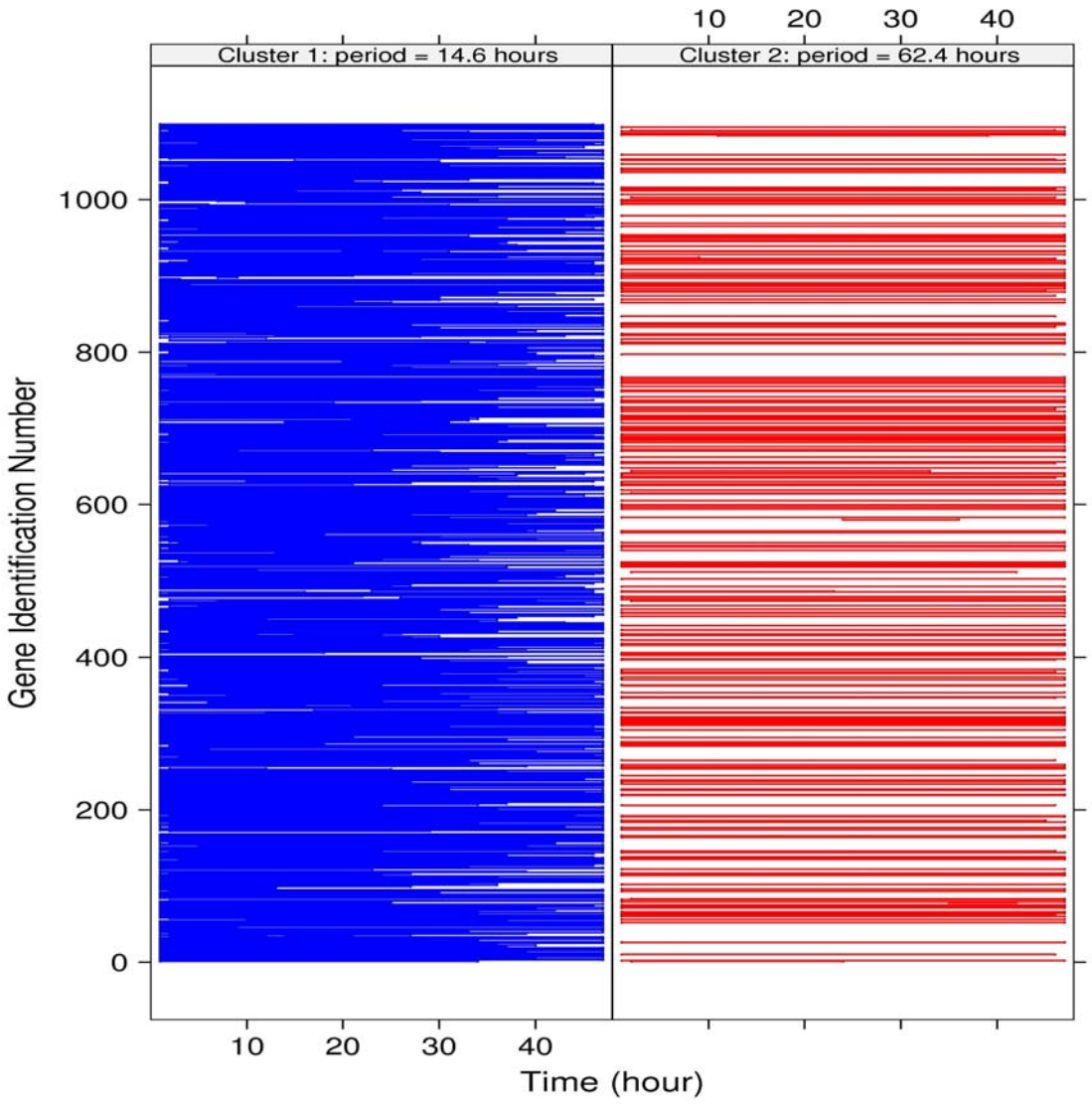



**Figure 7.** Two dynamic clusters for the HeLa data (Whitfield et al. 2002). Left: the cluster with period of 14.6 hours. Right: the cluster with period of 62.4 hours. The segments represent genes, with starting and ending points as indicated. The majority of genes in the left cluster appear only in a partial time interval while the majority of segments in the right cluster remain across the whole time interval.

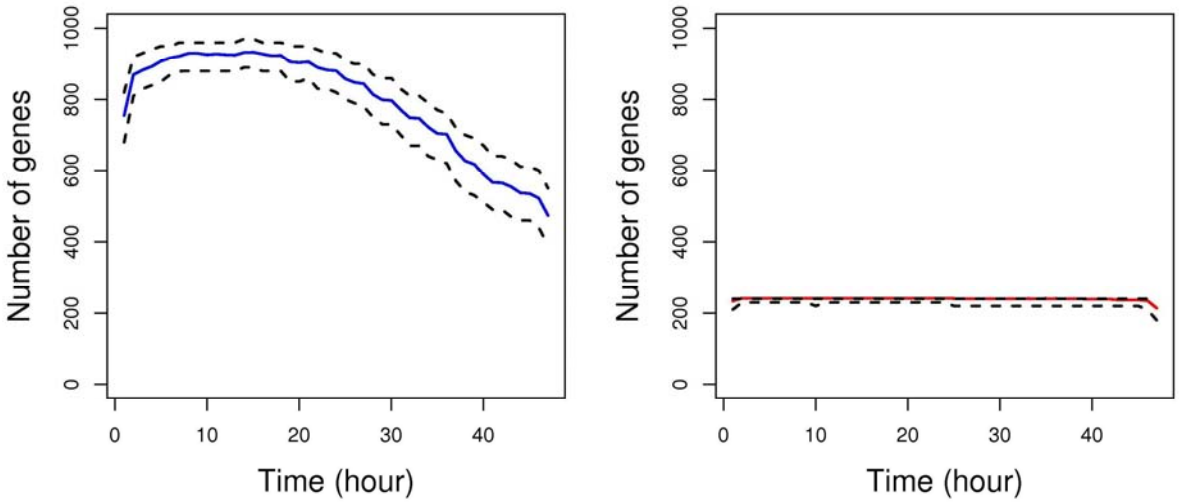



**Figure 8.** The number of genes over time in each cluster for the Whitfield et al. (2002) data. Left: the cluster with period 14.6 hours. Right: the cluster with period 64.2 hours. The black dashed curves are the 95% confidence band on the number of genes as obtained by the bootstrap method.